基于注意力网络的情感分析中的对比句处理 *

张 蓉 [1, 2†]，刘 渊 [2]，李 阳 [1]

(1. 江苏信息职业技术学院 物联网工程学院，江苏 无锡 214000; 2. 江南大学 人工智能与计算机学院，江苏 无锡 214000)

摘 要：方面级情感分析旨在确定评论中对特定方面的情绪极性，但目前较少研究复杂句对情感分类的影响。基于此，提出了一种基于 BERT 和带相对位置自注意力网络的方面级情感分析模型。首先，通过动态加权采样方法平衡对比句稀缺的问题，使模型学习到更多的对比句特征信息；其次，利用双头自注意力网络提取带相对位置的特征表示，与预训练模型得到的带绝对位置的特征表示联合训练；最后，通过标签平衡技术对模型正则化处理，稳定模型对中性样本的辨识。该模型在 SemEval 2014 Task 4 Sub Task 2 上进行实验，在两个数据集上的 Accuracy 和 Macro-f1 指标都有所提高。实验结果表明，该模型在对比句分类上是有效的，同时在整个测试集上分类也优于其他基准模型。

关键词：方面级情感分析；对比句；注意力网络；BERT 模型；相对位置编码

中图分类号：TP391 doi: 10.19734/j.issn.1001-3695.2022.02.0052

Handling contrastive sentences in sentiment analysis with attention network

Zhang Rong[1, 2†], Liu Yuan[2], Li Yang[1]

(1. *School of Internet of Things Engineering*, *JiangSu Vocational College of Information Technology*, *WuXi Jiangsu* 214000, *China*; 2. *School of Artificial Intelligence & Computer*, *JiangNan University*, *WuXi Jiangsu* 214000, *China*)

Abstract: Aspect-level sentiment analysis aims to determine the sentiment polarity towards specific aspect in reviews. However, little research has been done on the influence of complex sentences on sentiment classification. Based on this, this paper proposed an aspect sentiment classification model based on Bert and Self-attention network with relative position. Firstly, it used the dynamic weighted sampling method to balance the rare contrastive sentences, so that the model can learn more contrastive sentence feature information. Then, it jointly trained the feature representations extracted by double-head self- attention network with relative position and the feature representations obtained by the Pre-trained model with absolute position . Finally, it used the label smoothing regularization technology to stabilize the model to identify the neutral samples. It tested this model on Sub Task 2 in SemEval 2014 task, and improved both accuracy and Macro-F1 indicators of the two datasets. The experimental results show that the effectiveness of the proposed model for contrastive sentences classification, and also yield improvements in the whole test set over other benchmark models.

Key words: aspect-level sentiment analysis; contrastive sentences; attention network; bert model; relative position encoding

0 引言

文本情感分析可以帮助企业准确的分析用户对产品各个方面的评价，为企业制定详细产品更新策略提供有效参考意见。而且细粒度的情感分析方法，因其在对话系统、在线评论和社交网络等现实场景中的广泛应用而受到学术界和业界的关注和兴趣[1]。方面级情感分类(aspect-level sentiment classification，ASC)[2]是一种细粒度的情感分析任务，它旨在确定一个句子中方面词的情感极性(例如：积极、消极、中性)。一个句子可能含有多个方面词，且每个方面词的情感极性可能不同，所以需要指定目标方面才能判断相应的情感极性。

近年来，深度学习通过构建神经网络自动进行学习提取特征，在情感分析领域展现出良好的性能[3]。基于 BERT[4]等的预训练模型(Pre-trained Models，PTMs)就是其中最先进的模型之一，已被证明在 GLUE 基准测试[5]上具有最先进的性能，包括文本分类。BERT 是一个在维基百科大型文本语料库中预先训练过的语言模型，它特殊的结构允许对有监督的 ASC 等下游任务进行微调。虽然 PTMs 从大型语料库中获取一般语言知识，本身已经包含了很丰富的语法语义知识，但如何有效地将其知识适应下游任务仍然是一个关键问题[6]。同时，由于维基百科文章内容客观陈述的多，而带情感的主观评论少，导致 BERT 模型对情感方面的内容学习不够；再加上通常用于 ASC 分析的数据集都只有少量的训练样本，使得原本就复杂的 ASC 任务依然面临着严峻的挑战[7]。

另外，ASC 任务不仅缺乏带有标签的训练数据，还存在复杂句问题(例如对比情绪句、内隐情绪句和具有误导性的中性评论)，如表 1 所示，例如评论“air has higher resolution but the fonts are small.”，这个句子就存在两个个目标方面：“resolution”“fonts”，和相互对立的情绪极性：“higher”为积极，“small”为消极；在评论“The waiter poured water on my hand and walked away.”中并不包含情感词，但面向目标方面“waiter”很明显呈现消极；而评论“The service was typical short-order, dinner type.”表述非常隐晦，面向目标方面“service”情感极性为中性。以上这些复杂句都超出了现有模型的学习能力[8]。

该文通过进一步研究 BERT 作为预训练模型的不足和

收稿日期：2022-02-17；修回日期：2022-04-16 基金项目：国家自然科学基金资助项目(61972182)；江苏省高等学校自然科学研究面上项目基金资助项目(18KJD510011)；江苏省高等职业教育高水平专业群建设项目资助项目(苏教职函〔2021〕1 号)；江苏省高等职业教育产教融合集成平台建设计划项目资助项目(苏教职函〔2019〕26 号)

作者简介：张蓉(1980-)，女(通信作者)，副教授，访问学者，硕士研究生，主要研究方向为自然语言处理(95291188@qq.com)；刘渊(1967-)，男，教授，博导，主要研究方向为数字媒体、网络安全；李阳(1990-)，男，讲师，博士研究生，主要研究方向为图像处理与计算机视觉.

ASC 数据集中复杂句(例如，不同方面具有不同极性的句子)的分布与特征，基于 BERT-DK[7]，在采样、特征提取等方面优化与复杂句特征相关的微调技术。该文的主要贡献有：1) 对验证集上的错误样本进行实证研究，系统总结易错样本的特征；2) 提出了一种新的方面级情感分析框架，提高了复杂句和整个测试样本的分类性能；3) 将利用注意力模块提取出带相对位置的特征表示，与 BERT-DK 模块提取出带绝对位置的特征表示联合训练，提升模型对位置信息的捕获能力；4) 首次将加权随机采样应用于方面级情感分析中。

表 1　评论中复杂句示例

Tab. 1　Examples of complex sentences in reviews

评论	复杂句类型
air has higher reso- lution but the fonts are small.	对比句
The waiter poured water on my hand and walked away.	内隐句
The service was typical short-order, dinner type.	误导性中性句

1　相关工作

1.1　方面级情感分析

ASC 任务既可以单独训练，也可以和方面词提取(Aspect Extraction, AE) 任务一起联合训练[9, 10]。它需要关注每个具体方面的细微观点，因此相较于篇章级或句子级的情感分类任务更加复杂[11]。其中 ASC 任务的小样本问题一直受到研究者的重视，通常采取两种解决途径：一种是通过模型优化，使其更擅于捕获语法、语义特征，例如：Sun Chi 等人[12]通过对方面构造辅助句的方式将 ASC 任务从单个句子分类任务转换为语句对的分类任务，类似于机器问答和自然语言推理任务，通过对 BERT 预训练模型微调获得更佳性能。Karimi 等人[13]利用对抗过程在嵌入空间中生成与真实世界的例子类似的数据，对情绪分析中的 AE 和 ASC 两个任务联合对抗训练，提出了 BERT 对抗训练(BAT)的新架构。另一种是通过辅助额外的情感字典或同领域语料库，例如：He Ruidan 等人[14]提出的 PRET+MULT 框架通过共享浅层嵌入和 LSTM 层的方式从亚马逊评论数据集上训练的文档级情绪分类任务的情感知识迁移到 ASC 任务。文献[7]通过使用额外特定领域的数据，提出一种后训练方法微调 BERT 模型从其源领域和任务适应到方面级情感分析领域和任务中。

然而，最近基于神经网络的方法较少关注 ASC 数据集中包含复杂句的问题。Xu Hu 等人[15]用具体数据和实验结果证实 ASC 数据集中存在对比句且极其罕见(有多个方面且具有不同极性的句子称为对比句)，导致现有的 ASC 分类器不能很好地学习这些对比句知识，从而“降级”为句子级的情感分类器。并提出了一种自适应重加权(ARW)方案，通过给每个训练样本分配一个代表训练重要性的权重，动态地将模型引导向强化对比句的样本训练上，有效地提高了对比句的样本分类。Li Zhengyan 等人[16]对内隐情绪句进行专门的研究，将 ASC 数据集划分为外显式情绪表达切片和隐式情绪表达切片，结果表明大约有 30%的评论被划分为隐式情绪表达。通过在大规模情感注释的语料库上采用监督对比预训练引入外部情感知识来，将内隐情绪表达的表现与具有相同情绪标签的表现对齐，并采用方面感知微调来提高模型对基于方面的情绪识别的能力。

1.2　调整样本权重

大多数用于分类的机器学习算法都是在假设平衡类的情况下开发的，然而，在现实生活中，拥有适当平衡的数据并不常见。在自然语言处理任务中，也存在大量的类别不平衡的任务。最经典的就是序列标注任务中类别是严重不平衡的[17]，比如在命名实体识别中，显然一句话里边实体是比非实体要少得多，这就是一个类别严重不平衡的情况。对于缓解类别不平衡问题，比较基本的方法就是调节样本权重[18]。在学习中为少数类样本赋予更高的权重，比如在神经网络中，使得少数类产生的误差损失对网络权重更新贡献更大。调整样本权重在领域适应[19]和情感分析[20]都得到应用，但加权目的和加权方法完全不同。该文则是通过调整对比句采样权重来改善罕见但关键的样本在训练中的影响性。

1.3　自注意力机制

自注意机制(Self-Attention)[21]自提出以来，迅速在自然语言处理领域取得了巨大的进展。对于文本分类和推荐等任务，输入是一个序列，但输出不是一个序列，在这种情况下，注意力可以用于学习相同输入序列中的每个 token 的与之相关的 token，注意力权重的目的是捕捉同一序列中的两个单词是如何关联的，其中相关性的概念取决于主要任务[22]。

对于给定的词嵌入层输出序列 $X=(x_1,\ x_2,\dots,x_n)$，为每个序列位置创建三个向量(Q-查询向量、K-键向量、V-值向量)，然后对每个位置 x_i 使用 Q、K、V 实现注意力机制，最终得到序列 $Y=(y_1,y_2,\dots y_n)$，其中 y_i 包含了 x_i 的信息以及 x_i 与所有其他序列位置的关系。这里的 Q、K、V 三个向量使用前馈层生成。自注意力计算公式为

$$\text{Attention(Q,K,V)}=\text{softmax}\left(\frac{\text{QK}^{\text{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

对于给定词向量维度 d_{model}，多头自注意力即指对维度为 d_{model}/h 的投影(Q、K、V)矩阵执行 h 次注意。对每一个 head，(Q、K、V)被唯一的投影为维度为 d_{model}/h 的矩阵，自注意力输出维度也为 d_{model}/h。然后将每个 head 的输出连接起来，并再次应用线性投影层，得到与在原始(Q、K、V)矩阵上执行一次自注意相同维度的输出。整个过程用公式描述如下：

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{head}_1,\dots\text{head}_h)W^o \tag{2}$$

$$\text{head}_i=\text{Attention}(QW_i^Q,KW_i^K,VW_i^V) \tag{3}$$

其中投影为权重矩阵 $W_i^Q\in\mathbb{R}^{d_{model}\times d_q}$，$W_i^K\in\mathbb{R}^{d_{model}\times d_k}$，$W_i^V\in\mathbb{R}^{d_{model}\times d_v}$，$W^o\in\mathbb{R}^{hd_v\times d_{model}}$。

在设计 ASC“预训练+微调”架构分类器时，自注意力机制经常被运用于对下游任务的微调中[23]。实际上 BERT 等预训练模型的主体是 Transformer，而 Transformer 有两个主要组成部分：自我注意和位置级前馈层。两者都是排列等变的，并且对输入标记的顺序不敏感。为了使模型具有位置感知性，会通过自注意力机制在每个字符位置的词嵌入的基础上，添加绝对位置编码，但是这种绝对位置编码方式会导致一些片段位置信息损失[24]。

受 Shaw 等人[25]的启发，该文提出了 DWS+ RpSAN 来解决 BERT 预训练模型位置信息损失和 ASC 任务中复杂句的挑战。和以往工作的主要区别首先是利用将对比句稀缺问题看做简单的不平衡类问题，通过动态调整训练样本的采样权重来提升对比句的采样频率；其次结合了 BERT 预训练模型提取的绝对地址特征和自注意力模块提取的相对地址特征，通过并行训练，弥补了预训练模型在位置信息提取方面的不足。

2　数据集分析

该文在当下最受欢迎的 ABSA 基准数据集 SemEval 2014 Task 4 Sub Task 2 [26]上评估算法的性能，该数据集包括 2 个方面的领域：餐厅(简称 Rest)和笔记本电脑(简称 Lap)，拥有 3 种情绪标签：积极、消极和中性，每个评论包含 0 个、1 个或多个目标方面，这些评论包含不规则的词汇单位和句法模式，因此，这些数据是有噪声的、稀疏的和高维的。为了与前人的实验结果进行比较，确保实验数据的一致性，该文采用了 Xu Hu 等人[15]实验中对原始数据集的处理，包括删除了原始数据集中存在冲突的句子，对每条评论添加“contra”标

签，并从测试集中抽取出对比性句子创设一个独立的数据集以测试比较各模型处理性能。详细统计数据如表2和3所示。

Pontiki[26] 等人在对这两个领域数据集注释过程中发现：在 Lap 数据集中，用户对笔记本电脑的评论多为一个整体，并且当他们对特定方面评论时，通常使用形容词隐含指代某些方面(例如“昂贵”、“重”等)，而不使用具体明确的目标方面(例如“价格”、“重量”等)，因此相比之下 Rest 数据集包含了更多的目标方面。由表 2 可见，在训练集中，含有方面的句子占比在 Rest 数据集中达到 75%，而在 Lap 数据集中仅为 47.75%。此外，对笔记本电脑的评论经常提到功能描述而不表达任何情绪(例如“Has a 5-6 hour battery life.”)，从而导致 Lap 数据集中包含了更多的中性样本。

表 2　SemEval 2014 Task 4 Sub Task 2 方面情绪分类统计表

Tab. 2　Summary of semeval 2014 Task 4 Sub Task 2 on aspect sentiment classification

	Rest	Lap
训练集		
#句子	2000	3045
#方面	1743	2358
#积极	2164	987
#消极	805	866
#中性	633	460
# 含有方面的句子	1978	1462
%含有方面的句子	75%	47.75%
# 对比性句子	319	165
%对比性句子	16.1%	11.3%
测试集		
#句子	676	800
#方面	622	654
#积极	728	341
#消极	196	128
#中性	196	169
# 含有方面的句子	600	411
%含有方面的句子	88.8%	51.4%
# 对比性句子	80	38
%对比性句子	13.3%	9.2%

另外，最能体现和评估模型处理细粒度情感分类性能的对比性情绪句在这两个领域的训练集和测试集中都很罕见。Rest 训练集上的对比性句约 16%，而 Lap 训练集上更少(约 11%)，甚至比注释错误的句子(可以看做是噪声)还要少。在这样一个对比句匮乏的数据集上训练的机器学习模型倾向于降级为粗粒度(句子级)情感分类器。例如，对于评论“The screen is good and also the battery.”尽管存在两个目标方面“screen”和“battery”，但每个目标方面情感极性都为积极，使得整句也呈积极，模型处理这样的评论相当于处理句子级的情感分类。事实上，大多数样本主导训练过程，罕见但重要的样本很容易被忽略，甚至可能被认为是噪声，这对于大多数机器学习模型来说是一个普遍且广泛存在的问题，可以看做不平衡数据问题。

表 3　对比性句子测试集统计表

Tab. 3　Summary of Contrastive Test Set

对比性句子测试集	Rest	Lap
#对比性句子	80	78
#方面	228	203
#积极	85	72
#消极	60	71
#中性	83	60

通过上述分析，可以看出两个领域的数据集不仅存在小样本问题，还因为对某一特定主题发表意见的评论通常整句呈现出一种一致性意见而非对比性意见，在这种情况下，存在类数据不平衡问题，任何分类器都会偏向于多数非对比句，而这些问题在 Lap 数据集上尤为突出。

3　提出的方法

3.1　问题定义

给定一个上下文序列 $W^c=\{w_1^c,w_2^c...,w_n^c\}$ 和 $W^t=\{w_1^t,w_2^t...,w_n^t\}$，其中 W^t 是 W^c 的子序列，方面级情感分析任务旨在预测目标方面 W^t 在句子 W^c 中的情感倾向。图 1 展示了所提出的相对位置自注意编码器网络(DWS+RPSAN)的整体架构，它主要由融合领域意识的 BERT- DK 嵌入层、带相对位置自注意力编码层和输出层构成。

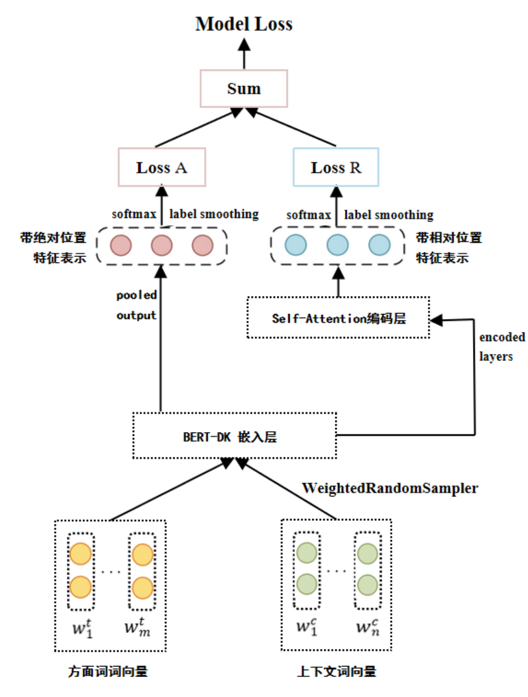


图 1　BERT-DK+DWS+RPSAN 模型结构

Fig. 1　Bert-dk+dws+rpsan model architecture

3.2　动态加权随机采样

鉴于对比句是细粒度情感分析任务的关键样本又非常罕见，需要思考现有的训练过程中如何使得机器学习模型从这些罕见的样本中学习。假设将数据集中样本分为对比性和非对比性两类：$class_{contra}$ 和 $class_{no_contra}$，基于均匀分布，从每个类中随机采样得到的概率为

$$p(x\in class_i)=\frac{\#\{class_i\}}{\#\{train\}}=\frac{N_{class_i}}{N_{train}} \tag{4}$$

而 Rest 训练集上 $class_{no_contra}:class_{contra}\approx 5$，Lap 训练集上 $class_{no_contra}:class_{contra}\approx 9$，两个数据集上 $N_{class_{no_contra}}\gg N_{class_{contra}}$，即 $p(x\in class_{no_contra})\gg p(x\in class_{contra})$，如果使用这样的数据集训练模型，那么模型看到的非对比性句子机会要远大于对比性句子，导致深度学习模型很难从现有的训练过程中学习这些罕见样本。Gao 等人[27]通过研究发现，在训练的早期阶段，大多数样本的损失主导了总损失，并决定了模型参数更新方向。到了迭代后期，尽管罕见样本主导总损失，但是可能对总损失贡献不足。在最坏的情况下，当优化器开始过拟合大多数样本中的小细节时，才可能会考虑到罕见的例子的损失，意味着验证过程中为了避免过拟合可能会在罕见样本真正得到良好优化之前停止对模型的训练。

考虑到这种罕见但重要的样本很容易被忽略的机器学习过程，需要解决两个问题：一是在训练早期阶段增加对比句样本；二是在验证过程找到最佳模型之前更早地增加(或重新

平衡)那些没有被很好优化的样本的采样机会。一个自然解决方案是平衡训练集，多数类过采样或少数类过采样是两种可能的策略。由于数据非常稀疏，欠采样的多数类是次最优的，因为可能会在学习过程中失去有意义的样本。因此，过采样少数类是一个更好的解决方案[27]。在采样时对罕见且重要样本进行数据加强，使得 $p(x \in class_{no_contra}) = p(x \in class_{contra})$。由于深度学习模型通常是在逐批处理的基础上进行训练的，调整每类样本的权重自然应该是在每轮迭代结束时，这是因为每个样本都参与学习过一次，模型可以集中于那些没有被很好地处理(分类错误)的样本。

基于上述分析，该文的目标是设计一个动态自适应方案，不断调整在训练集中已知的对比句采样的权重。由于概率的数值不能明确区分模型是否在一个样本上犯了错误，所以实验中使用正确率加权法，即通过控制权重赋予对比性句更大的采样权重。同时，为避免模型过度适应少数类，每轮迭代结束后，找到分类不正确的样本和对应的类别，根据验证集上两种类别样本分类错误率动态更新权重。设第 n 轮迭代 $class_i$ 类样本的采样权重为 $w_i^{epoch_n}(i \in (no_contra, contra))$，初始化权重设置如下：

$$w_i^{epoch_n} = \frac{total_sample - num_{class_i}}{total_sample} \qquad n=1 \tag{5}$$

权重更新公式如下：

$$w_i^{epoch_n} = w^{epoch_{n-1}} + \varepsilon \times error_rate_i^{epoch_{n-1}} \qquad n \ge 2 \tag{6}$$

其中 $error_rate_i^{epoch_{n-1}}$ 表示第 n-1 次 epoch 中验证集上 $class_i$ 类的分类错误率，ε 是更新因子，用于调节验证集中分类错误样本类别对加权采样的影响，能影响下一轮迭代中对比句采样权重，越大则分类错误样本的类别对下一轮迭代加权采样影响越大，反之，对下一轮迭代加权采样的影响就越小。该文尝试 $\varepsilon \in$ {0.0、0.05、0.1、0.15、0.2、0.25、0.3}，ε=0.0 时，相当于每轮迭代时的各类别采样权重为初始化权重保持不变，此时模型在对比句分类上已获得了一定的性能提升，随着 ε 值的提升，模型性能并不是一直提升，而是先递增后递减，在 ε=0.1 时获得最好结果，这也说明 ε 值过大存在过分强调分类结果对采用权重的影响。

3.3　BERT-DK 嵌入层

词嵌入层使用预先训练好的 BERT-DK 模型[6]来生成序列的词向量,该模型是在 BERT[3]的基础上，针对 BERT 对情感方面的内容学习不够，不能很好适用于评论类分类特别是细粒度情感分类的问题，先进行掩码语言建模，然后使用无监督的领域(餐厅或笔记本电脑)评论数据集对预先训练过的 BERT 权重进行句子预测，提高其领域意识，再使用有监督的 ASC 数据进行微调。经过 BERT-DK 层处理过的词向量从某种程度上说是具备领域意识的。

3.4　带相对位置的自注意力层

注意力机制属于非递归模型，无法捕捉输入序列中元素的顺序，因此使用时需要显式地编码位置信息。目前常用的有三种嵌入方式：正弦位置编码、通过学习得到的位置编码和相对位置表示。BERT-DK 嵌入层模块已使用正弦位置信号嵌入绝对位置信息，用于处理序列问题，而相对位置信息在执行自注意力计算时是丢失的，会导致和微调的实际数据之间存在偏差。为了加入这丢失的相对位置信息，该文使用 Shaw 等人[25]提出的相对位置嵌入。相对位置嵌入不是对每个位置使用固定的嵌入，而是根据自我注意机制中比较的"键"和"查询"之间的偏移量产生不同的学习嵌入。在原自注意力基础上，引入了两个只与相对位置有关的向量：$a_{ij}^V, a_{ij}^V \in \mathbb{R}^{d_{model}}$，学习每两个序列位置之间的相对位置信息，其采用一组可训练的嵌入向量来表示输入句子中每个单词的位置编码。如果 attention 的目标词是 x_i 的话，引入这两个向量之后，那么在计算 x_j 对 x_i 的注意力特征的时候，需要额外考虑 x_j 对 x_i 的两个与位置相关的向量。同时引入一个可调参数 k，用于限制两个序列位置之间最大的距离。实验中尝试了 $k \in \{1,2...,12\}$，发现 $k > 8$

算法 1　BERT-DK+DWS+RPSAN

Algorithm 1: BERT-DK+DWS+RPSAN

Input : D_{tr} : training set with n samples;

e: maximum number of epochs.

Output: $p_\theta(\hat{y} | \cdot, \cdot)$: a trained model.

$w_i \leftarrow \frac{n - n_{class_i}}{n} \qquad i \in (no_contra, contra)$

// 初始化所有样本权重

for $epoch \in \{1, ..., e\}$

do

for $(ab, xb, yb) \in Batchify(D_{tr}, w_i)$

// 检索一个动态加权随机采用的批次

do

$emb_{encoder_layers}, emb_{pooled} \leftarrow Bert\text{-}DK(a^b, x^b)$

// 使用 bert-dk 模型预训练

$emb_{rp} \leftarrow RPSan(emb_{encoder_layers})$

// 使用带相对位置的自注意力层提取特征

$L_{Absolute} \leftarrow CrossEntropy(p_\theta(\hat{y}^b | emb_{pooled}), y^b)$

// 计算带绝对地址词嵌入的交叉熵损失

$L_{Relative} \leftarrow CrossEntropy(p_\theta(\hat{y}^b | emb_{rp}), y^b)$

// 计算带相对位置词嵌入的交叉熵损失

$L(\theta) = \frac{L_{Absolute} + L_{Relative}}{2} + L_{lsr}$

// 计算带标签平滑正则化的联合训练损失

$\text{BackProp\& ParamUpdate}(L, M)$

// 反向传播并更新参数

end

$\hat{y}_{1:n} \leftarrow \text{argmax}\, p_\theta(\hat{y}_{1:n} | a_{1:n}, x_{1:n})$

// 计算当前预测结果

$error_rate_i \leftarrow \frac{\sum_1^j (y_j \ne \hat{y}_j \Lambda class_i(x_j))}{n}$

// 计算错误率

$w_i = w_i + \varepsilon \times error_rate_i$

// 调整所有样本权重

end

以后效果就没有提升了，说明临域为 8 的窗口内，attention 对相对位置比较敏感，窗口以外，相对位置可以不做区分，即临域为 8 的窗口以外，用窗口内(gram=8)的嵌入向量平均池化替代。将 a_{ij}^V, a_{ij}^V 定义为可训练的向量，本质上就是训练 $w^K = (w_{-k}^K, ..., w_k^K)$ 和 $w^V = (w_{-k}^V, ..., w_k^V)$：

$$a_{ij}^K = w_{clip(j-i,k)}^K \tag{7}$$

$$a_{ij}^V = w_{clip(j-i,k)}^V \tag{8}$$

$$\text{clip}(x,k) = \max(-k, \min(k,x)) \tag{9}$$

3.5　分类模型

由于现有 ASC 模型对中性情绪分类性能不够稳定，主要表现在一方面易于将中性样本分类成积极或消极，另一方面易于将积极或消极样本分类成中性极性，BERT-DK+DWS+RPSAN 模型在原有的交叉熵损失函数中引入一个标签平滑正则化 (Label Smoothing Regu- larization，LSR)项[28]，用以惩罚低熵输出分布，抑制模型对其预测的自信度，达到稳定模型对中性类的判断的目的。对于训练样本 $\boldsymbol{x}$，假设其每个标签 $k \in \{1...k\}$ 的实际概率分布为 $q(k|x)$，将 $q(k|x)$ 替换为

$$q(k|x) = (1-\lambda)q(k|x) + \lambda u(k) \tag{10}$$

其中 u(k)是标签上的先验概率分布，λ 为平滑参数。实验中，

先验标签分布统一设置为 $u(k)=1/C$ 。

LSR 等价于标签先验概率分布 $u(k)$ 与模型预测分布 p_θ 之间的 KL 散度。LSR 定义如下：

$$L_{lsr}=-D_{KL}(u(k)\|p_\theta) \tag{11}$$

因此，LSR 相当于用一对交叉熵损失 $q(k|x)$ 和 $u(k)$ 替换一个交叉熵损失。

整个模型需要优化的目标函数(损失函数)是 $L_{Absolute}$ 、$L_{Relative}$ 和 L_{lsr} 的交叉熵损失，其定义为

$$L(\theta)=\frac{L_{Absolute}+L_{Relative}}{2}+L_{lsr} \tag{12}$$

其中，$L_{Absolute}$ 和 $L_{Relative}$ 分别代表带绝对地址特征表示和带相对地址特征表示的损失。

BERT-DK+DWS+RPSAN 模型的参数优化过程如算法 1 BERT-DK+DWS+RPSAN 所示，提出的算法包含三个阶段：用动态加权随机采样阶段，用 BERT-DK 词嵌入预处理阶段，联合带相对位置自注意力机制微调阶段。

4　实验及分析

4.1　实验环境和超参数设置

所有的实验和基准测试都使用一个单一的 GPU(GTX 1080 Ti)运行，CPU 为 Intel Core i7-8700K@4.7 GHz，内存为 16G。

微调时，对于超参数的设置一般与所参考和对照的实验保持一致，个别超参数也会根据新模型特点进行调整。其中，自注意力模块中多头注意力 head 的个数与 Shaw 等人[24]所设置的一致，将文本向量分成 2 个头效果是最好的。在辍学率上，与前者的高辍学率 0.7 不同，更倾向于 0.1 这样的低辍学率，这与本模型基于 BERT 有关，一般 BERT 模型在处理情感分类时，都会选择 0.1 作为辍学率。在 Rest 和 Lap 数据集的初始学习率选择上，2e-5 和 3e-5 都经过前人实验反复验证过的比较好的设置值，该文通过设计多种消融反复实验，发现将 Rest 数据集初始学习率设置为 2e-5，将 Lap 数据集初始学习率设置为 3e-5 是最合适的，这种细微的差别应该和 Rest 数据集相比 Lap 数据集包含更多的方面级句子有关。批量大小设置为 32，与 BERT-DK 模型保持一致，每批都是通过加权随机采样训练集构建。在训练过程中，设置 epoch 数为 20，并保存在此期间训练得到的最大准确率模型。带相对位置的自注意力模型的词向量维度 d_{model} 与 BERT-DK 模型输出词嵌入相同，设置为 300；采用 Adam 优化器对所有参数进行更新，设置交叉熵损失的标签平滑参数 λ 为 0.2。所有结果的平均运行次数超过 10 次。

4.2　对比模型

该文选取 4 个分类器作为基线，同时对该文提出的合成新的改进模块进行了消融实验，结果表明所有的模块都对最后的性能有帮助作用，而其中动态随机加权采样对 Rest 数据集提供了最大的贡献，带相对位置的自注意力层对 Lap 数据集提供了最大的贡献。

AOA[29]：引入了一个 attention-over-atten- tion(AOA)神经网络，以联合的方式对目标方面和句子进行建模，并明确地捕捉方面与句子上下文之间的交互作用。

MGAN[30]：利用细粒度和粗粒度的注意机制来设计 MGAN 框架，还使用目标方面对齐损失来描述具有相同上下文的目标方面之间的方面级交互。

BERT-DK[7]:在 BERT 模型基础上，使用域(笔记本电脑或餐厅)评论，对预先训练过的 BERT 权重首先执行掩码语言(MLM)建模，再进行下一句子(NSP)预测，然后使用有监督的 ASC 数据进行微调。

4.3　实验结果与讨论

模型的性能采用 *Accuracy* 和 *Macro-f1* 度量来评估，表 4 包含了所有实验结果的总结。

表 4　各模型在完整数据集和对比句测试集上性能

Tab. 4　Performance of ASC baselines and the proposed Scheme on both Full Test Set and Contrastive Test Set

模型	Rest		Lap	
	Acc.	MF1	Acc.	MF1
AOA				
完整数据集	81.20	-	74.50	-
对比句测试集	42.98	33.66	42.86	33.53
MGAN				
完整数据集	81.25	71.94	75.39	72.47
对比句测试集	53.95	57.64	46.80	43.38
BERT-DK				
完整数据集	84.21	76.2	76.9	73.65
对比句测试集	65.53	66.92	51.13	50.04
BERT-DK+ARW				
完整数据集	85.35	78.46	77.23	73.81
对比句测试集	71.84	72.66	61.08	60.34
Ours:BERT-DK+DWS				
完整数据集	86.33	79.60	*76.39*	*72.95*
对比句测试集	73.90	74.91	62.81	61.97
Ours:BERT-DK+RPSAN				
完整数据集	86.09	79.42	77.82	74.89
对比句测试集	71.97	73.28	67.34	66.54
Ours:BERT-DK+DWS+RPSAN				
完整数据集	86.88	81.48	78.90	75.63
对比句测试集	75.90	75.78	67.00	66.40

从实验结果可以看到采用动态加权采样方法后，Rest 和 Lap 的对比句测试集性能分别提高了约 8.4%和 11%。与 BERT+DK 方案相比，也提高了约 2%。在加权采样后，Rest 完整数据集的性能在 Rest 上有所改善，但在 Lap 上的性能略有下降，原因可能是加权采样不适合学习，而 Lap 数据集中噪声样本(注释错误样本)远超对比句样本，该模型学习了更多的一些注释错误使得整体性能下降。

BERT-DK+RPSAN 模型使 Rest 和 Lap 的对比句测试集性能分别提高了约 6.4%和 15.5%，值得一提的是，在 Lap 的对比句测试集上性能达到了最优。与 BERT+DK 相比，在 Rest 对比句测试集上性能略有提升，在 Lap 对比句测试集上提高了约 4.5%。在 Rest 和 Lap 的完整数据集上，性能较 BERT-DK 和 BERT-DK+ AWS 比均有提升。

BERT-DK+DWS+RPSAN 模型除了在 Lap 的对比句测试集上性能略低于不带加权采样的 BERT-DK+RPSAN 模拟，其他性能上均有提升，在 Rest 的对比句测试集上尤为显著，证明了该文的改进方法是有效的，特别是提高了模型在对比句上的整体表现，真正测试了细粒度层面的情感分类能力。

图 2 展示了分别使用随机采样和动态加权随机采样方法从 Rest 测试集最后 10 个批次中采样的非对比性句子类和对比性句子类的分布情况，每个批次的左侧柱体代表非对比性句子类，右侧柱体代表对比性句子类，由图所示，加权随机采样很好平衡了数据集中关键性对比性句子严重稀缺的问题。

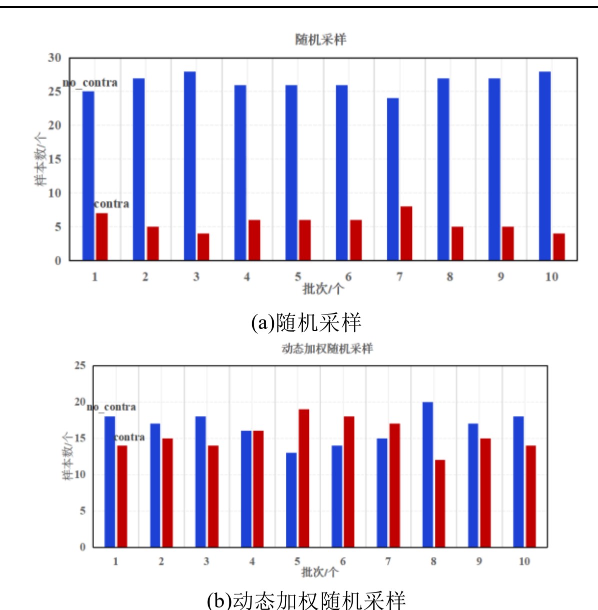
表 5 是运行 BERT-DK+DWS 运行在验证集上的易错样本分析统计表，运行 10 次分类错误次数超过 5 次的样本被定义为“Hard Sample”，分析这些样本，可以看出模型对中性分类最易出错(Rest 和 Lap 数据集中的 Hard Sample 均为 22 个，占比 70%以上)，具体体现倾向于把 中性标签预测成其他标签，或者是把其他标签预测成中性，主要原因可能在于中性情绪本就是一种非常模糊的情感状态，同时与数据集人工标注中性存在不可靠性有关；另外，模型对整句中存在对

比性意见的句子(含对比句)预测出错率占比高(Rest 数据集中 20 个,Lap 数据集中 18 个，占比 58%以上)，而且方面词在整句中所处的位置非常关键。

与数据集人工标注中性存在不可靠性有关；另外，模型对整句中存在对比性意见的句子(含对比句)预测出错率占比高(Rest 数据集中 20 个,Lap 数据集中 18 个，占比 58%以上)，而且方面词在整句中所处的位置非常关键。

此处用 2 个非对比句易错样本为例来展示中性分类不稳定性和整句中存在对比性意见时，方面词与代表情绪性词汇相对距离不同，分类难易不同，具体见图 3 和 4。

增加带相对位置的自注意力模块后，在 HardSample1 方面词[leather carrying case]的 10 次预测中，错误预测由 7 次降为 4 次，不再是易错样本，而 HardSample2 方面词[application]的 10 次预测中，错误预测由 10 次降为 8 次，但还是属于易错样本。一方面说明自注意力模块能够提升模型对中性类的分辨能力，模型泛化能力增强；另外一方面也说明中性类不稳定和整句中存在对比性意见对分类结果的影响等问题依然是 ASC 任务的瓶颈，这也是作者后面的研究重点。标签平滑正则化(LSR)的处理使得中性类样本的预测准确率得到一定的提升，从三个消融模型的实验结果来看，提升大概在 0.12-0.2%之间，表 4 中就不再详细标注。



(a)随机采样

(b)动态加权随机采样

图 2　Rest 测试集中最后 10 个批次采样的非对比句和对比句的分布情况

Fig. 2　Distribution of no_contra samples and contra samples in the last 10 batches of Rest test set

表 5　验证集上易错样本(Hard Sample)分析

Tab. 5　Hard sample analysis on validation set

数据集	总计	Hard Sample	Hard Sample 中与“中性”相关的样本		包含不同意见词汇的句子	
			样本极性为“中性”	预测极性为“中性”	对比句	非对比句
Rest	150	31	16	6	4	14
Lap	150	30	9	13	11	9

Hard Sample 1：

Neutral，10次预测，全部正确

第2个方面词

描述事实，不表达任何情绪

The leather carrying case, keyboard and mouse arrived in two days from Memphis warehouse.

第1个方面词

Neutral，10次预测，7次错误预测成Positive

第3个方面词

Neutral，10次预测，全部正确

整句呈Neutral

图 3　Hard Sample1 展示了中性类不稳定和相对位置对分类结果的影响

Fig. 3　Hard sample1 shows the neutral class' instability and the influence of relative position on classification results



图 4　Hard Sample2 展示了中性类不稳定和整句中存在对比性意见对分类结果的影响

Fig. 4　Hard sample2 shows the neutral class' instability and the impact of the existence of comparative opinions in the whole sentence on the classification results

5　结束语

本研究注意到 BERT 作为情感分类任务预训练模型时存在的位置信息损失问题，采用实证法对验证集上易错样本加以分析，证实了位置信息对分类的重要性。同时对 ASC 数据集中的复杂句进一步研究，观察到提升 ASC 分类器性能的关键不仅要解决对比句稀缺问题，还要注意中性类不稳定问题和整句中存在对比性意见时对其中具体方面正确分类所带来的挑战。该文通过动态加权采样方法平衡对比句和非对比句训练样本数量，并利用自注意力网络提取带相对位置的特征

表示和预训练模型提取的带绝对位置特征表示联合训练，辅以标签平滑正则化处理。实验结果表明，该模型在处理对 ASC 任务至关重要的对比句方面取得了新突破，同时在整个测试集上分类效果也很好。

参考文献：

[1] 张严，李天瑞. 面向评论的方面级情感分析综述 [J]. 计算机科学, 2020, 47 (6): 200-206. (Zhang Yan, Li Tianrui. Review of comment-oriented aspect-based sentiment analysis [J]. Computer Science, 2020, 47 (6): 200-206.)

[2] Thet T T, Na J C, Khoo C S. Aspect-based sentiment analysis of movie reviews on discussion boards [J]. Journal of Information Science, 2010, 36 (6): 823-848.

[3] Zhang Lei, Wang Shuai, Liu Bing. Deep learning for sentiment analysis: A survey [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8 (4): e1253.

[4] Devlin J, Chang Mingwei, Lee K, *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding [C]// Proc of the Conference of the North American Chapter of the Association for Computational Linguistics: Language Technologies. 2019: 4171-4186.

[5] Wang A, Singh A, Michael J, *et al.* GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding [C]// Proc of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. 2018: 353-355.

[6] Qiu Xipeng, Sun Tianxiang, Xu Yige, *et al.* Pre-trained models for natural language processing: A survey [J]. Science China Technological Sciences, 2020: 1-26.

[7] Xu Hu, Liu Bing, Shu Lei, *et al.* BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis [C]// Proc of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 2324-2335.

[8] Wang Kai, Shen Weizhou, Yang Yunyi, *et al.* Relational Graph Attention Network for Aspect-based Sentiment Analysis [C]// Proc of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 3229-3238.

[9] 曾义夫，蓝天，吴祖峰，等. 基于双记忆注意力的方面级别情感分类模型 [J]. 计算机学报, 2019, 42 (8): 1845-1857. (Zeng Yifu, Lan Tian, Wu Zufeng, *et al.* Bi-memory based attention model for aspect level sentiment classification [J]. Chinese Journal of Computers, 2019, 42 (8): 1845-1857.)

[10] Yang Heng, Zeng Biqing, Yang Jianhao, *et al.* A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction [J]. Neurocomputing, 2021, 419: 344-356.

[11] Ambartsoumian A, Popowich F. Self-attention: A better building block for sentiment analysis neural network classifiers [C]// Proc of the 9th Workshop on Computational Approaches to Subjectivity: Sentiment and Social Media Analysis. 2018: 130-139.

[12] Sun Chi, Huang Luyao, Qiu Xipeng. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence [C]// Proc of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 380-385.

[13] Karimi A, Rossi L, Prati A. Adversarial training for aspect-based sentiment analysis with bert [C]// the 25th International Conference on Pattern Recognition. IEEE, 2021: 8797-8803.

[14] He Ruidan, Lee W S, Ng H T, *et al.* Exploiting document knowledge for aspect-level sentiment classification [C]// Proc of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 579-585.

[15] Xu Hu, Liu Bing, Shu Lei, *et al.* A failure of aspect sentiment classifiers and an adaptive re-weighting solution [J]. arXiv preprint arXiv: 1911. 01460, 2019.

[16] Li Zhengyan, Zou Yicheng, Zhang Chong, *et al.* Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training [C]// Proce of the Conference on Empirical Methods in Natural Language Processing. 2021: 246-256.

[17] Akkasi A, Varoğlu E, Dimililer N. Balanced undersampling: a novel sentence-based undersampling method to improve recognition of named entities in chemical and biomedical text [J]. Applied Intelligence. 2018, 48 (8): 1965-1978.

[18] Guo X, Yin Y, Dong C, *et al.* On the class imbalance problem [C]// The 4th international conference on natural computation. IEEE, 2008, 4: 192-201.

[19] Wang Rui, Utiyama M, Liu Lemao, *et al.* Instance weighting for neural machine translation domain adaptation [C]// Proc of the Conference on Empirical Methods in Natural Language Processing. 2017: 1482-1488.

[20] Pappas N, Popescu-Belis A. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis [C]// Proc of theConference on Empirical Methods In Natural Language Processing. 2014: 455-466.

[21] Yang Zichao, Yang Diyi, Dyer C, *et al.* Hierarchical attention networks for document classification [C]// Proc of the conference of the North American chapter of the association for computational linguistics: human language technologies. 2016: 1480-1489.

[22] Chaudhari S, Mithal V, Polatkan G, *et al.* An attentive survey of attention models [J]. ACM Transactions on Intelligent Systems and Technology, 2021, 12 (5): 1-32.

[23] Yang Heng, Zeng Biqing, Yang Jianhao, *et al.* A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction [J]. Neurocomputing, 2021, 419: 344-356.

[24] Yang Zhilin, Dai Zihang, Yang Yiming, *et al.* Xlnet: Generalized autoregressive pretraining for language understanding [J]. Advances in neural information processing systems, 2019, 32.

[25] Shaw P, Uszkoreit J, Vaswani A. Self-attention with relative position representations [C]// Proc of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018: 464-468.

[26] Pontiki M, Galanis D, Papageorgiou H, *et al.* Semeval-2014 task 4: Aspect based sentiment analysis [C]// International workshop on semantic evaluation. 2014: 19-30.

[27] Gao T, Jojic V. Sample importance in training deep neural networks [J]. 2016.

[28] Szegedy C, Vanhoucke V, Ioffe S, *et al.* Rethinking the inception architecture for computer vision [C]// Proc of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.

[29] Huang Binxuan, Ou Yanglan, Carley K M. Aspect level sentiment classification with attention-over-attention neural networks [C]// International Conference on Social Computing: Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer, Cham, 2018: 197-206.

[30] Li Zheng, Wei Ying, Zhang Yu, *et al.* Exploiting coarse-to-fine task transfer for aspect-level sentiment classification [C]// Proc of the AAAI Conference on Artificial Intelligence. 2019, 33 (01): 4253-4260.